ICCV
#5746

A. Architecture Details

This section provides additional details about parts of the architecture not covered in the main paper.

AVCLIP preprocessing To align the resulting representation with the noised latent space z_t, we pre-processed the output of the AVCLIP encoder using two sequential transformation blocks. Each block consists of a fully connected (fc) layer, a ReLU activation, layer normalization, and dropout with $d = 0.1$. In the first block, the fully connected layer maintains the feature dimension at 768, while in the second block, it expands the feature dimension from 768 to 1024.

CLIP preprocessing Given a video V of 10 seconds at 25 FPS, we uniformly sampled 5 frames per second and computed their CLIP representations, resulting in feature vectors of shape $(1, 768)$ for each frame. We then applied linear interpolation, producing a signal of shape $(200, 768)$. To ensure alignment with the latent space, we symmetrically padded the signal on both sides so that the 10-second duration corresponds to 216 samples. Next, we processed the signal through a preprocessing block consisting of a fully connected (FC) layer, an ReLU activation, layer normalization, and dropout with $d = 0.1$. In this block, the FC layer preserves the feature dimension of the CLIP encoder output at 768. We then summed the result with the output of the first AVCLIP processing block and passed it through an additional preprocessing block, where the FC layer expands the feature dimension from 768 to 1024, ensuring that the final representation is aligned with the dimensions of z_t.

B. Training Details

Model	Steps/Epochs	Hardware	Batch Size	GPU Hours
CAFA	81k steps	A100 (40GB)	16	∼48
CAFA-base	48k steps	A100 (40GB)	16	∼24
MMAudio	300k steps	H100	-	304
FoleyCrafter	164+80 epochs	-	128	-
Frieren	2.4M steps	2 × RTX 4090	-	-
VATT	"3 days" of training	A100 (80GB)	-	∼72
MultiFoley	650k steps	-	128	-

Table 5. Comparison of training costs across different models. CAFA employs a two-stage training approach (VGGSound followed by VisualSound fine-tuning), while CAFA-base uses single-stage training. FoleyCrafter uses separate semantic (164 epochs) and temporal (80 epochs) training stages. Frieren employs a three-stage approach, and MultiFoley uses a two-stage training method.

C. User Study Form

Figure 5 shows the user study interface where participants compared audio outputs from different models, evaluating quality, temporal alignment with video, and adherence to textual prompts.

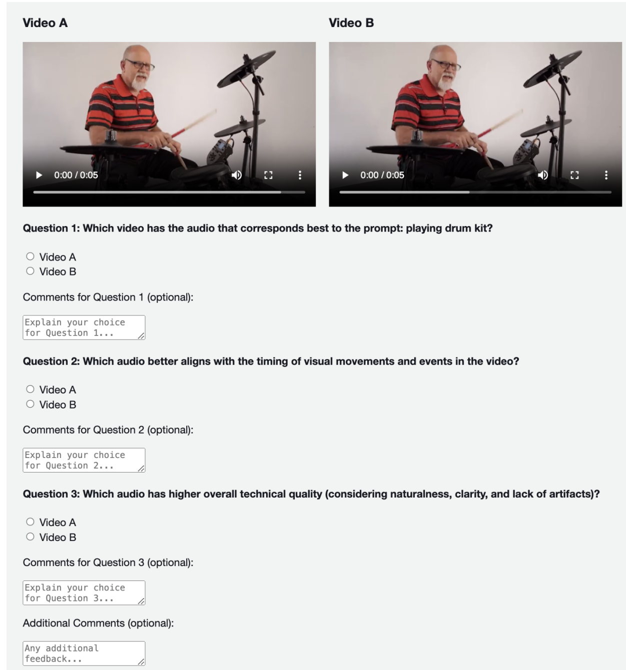

Figure 5. User study form

D. Additional Figures

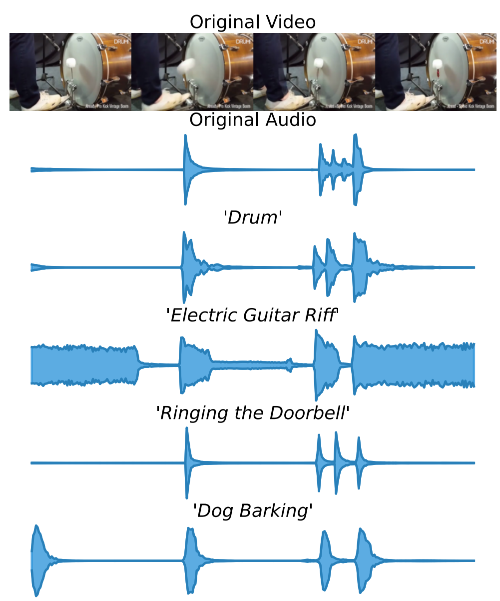

Figure 6. **CAFA creative control.** We demonstrate our method's ability to generate diverse, high-quality Foley sounds for videos through text prompts, ensuring temporal synchronization between audio and visual elements.